

SYSTEM AND METHOD FOR NOTATING POLYMERS

Related Application

5 This application claims priority under 35 U.S.C. §119 to US Provisional Patent Application Nos. 60/130,747, filed April 23, 1999, 60/130,792, filed April 23, 1999, 60/159,939, filed October 14, 1999, and 60/159,940, filed October 14, 1999, each of which are incorporated herein by reference in their entirety.

Background

10 Various notational systems have been used to encode classes of chemical units. In such systems, a unique code is assigned to each chemical unit in the class. For example, in a conventional notational system for encoding amino acids, a single letter of the alphabet is assigned to each known amino acid. A polymer of chemical units can be represented, using
15 such a notational system, as a set of codes corresponding to the chemical units. Such notational systems have been used to encode polymers, such as proteins, in a computer-readable format. A polymer that has been represented in a computer-readable format according to such a notational system can be processed by a computer.

Conventional notational schemes for representing chemical units have represented
20 the chemical units as characters (e.g., A, T, G, and C for nucleic acids), and have represented polymers of chemical units as sequences or sets of characters. Various operations may be performed on such a notational representation of a chemical unit or a polymer comprised of chemical units. For example, a user may search a database of chemical units for a query sequence of chemical units. The user typically provides a
25 character-based notational representation of the sequence in the form of a sequence of characters, which is compared against the character-based notational representations of sequences of chemical units stored in the database. Character-based searching algorithms, however, are typically slow because such algorithms search by comparing individual characters in the query sequence against individual characters in the sequences of chemical
30 units stored in the database. The speed of such algorithms is therefore related to the length of the query sequence, resulting in particularly poor performance for long query sequences.

Summary

In one aspect, the invention is directed to a notational system for representing
35 polymers of chemical units. The notational system is referred to as Property encoded

nomenclature (PEN). According to one embodiment of the notational system, a polymer is assigned an identifier that includes information about properties of the polymer. For example, in one embodiment, properties of a disaccharide are each assigned a binary value, and an identifier for the disaccharide includes the binary values assigned to the properties of the disaccharide. In one embodiment, the identifier is capable of being expressed as a number, such as a single hexadecimal digit. The identifier may be stored in a computer readable medium, such as in a data unit (e.g., a record or a table entry) of a polymer database. Polymer identifiers may be used in a number of ways. For example, the identifiers may be used to determine whether properties of a query sequence of chemical units match properties of a polymer of chemical units. One application of such matching is to quickly search a polymer database for a particular polymer of interest or for a polymer or polymers having specified properties.

In one aspect, the invention is directed to a data structure, tangibly embodied in a computer-readable medium, representing a polymer of chemical units. In another aspect, the invention is directed to a computer-implemented method for generating such a data structure. The data structure may include an identifier that may include one or more fields for storing values corresponding to properties of the polymer. At least one field may be a non-character-based field. Each field may be capable of storing a binary value. The identifier may be a numerical identifier, such as a number that is representable as a single-digit hexadecimal number.

The polymer may be any of a variety of polymers. For example, (1) the polymer may be a polysaccharide and the chemical units may be saccharides; (2) the polymer may be a nucleic acid and the chemical units may be nucleotides; or (3) the polymer may be a polypeptide and the chemical units may be amino acids.

The properties may be properties of the chemical units in the polymer. For example, the properties may include charges of chemical units in the polymer, identities of chemical units in the polymer, confirmations of chemical units in the polymer, or identities of substituents of chemical units in the polymer. The properties may be properties of the polymer that are not properties of any individual chemical unit within the polymer. Example properties include a total charge of the polymer, a total number of sulfates of the polymer, a dye-binding of the polymer, a mass of the polymer, compositional ratios of substituents, compositional ratios of iduronic versus glucuronic, enzymatic sensitivity, degree of sulfation, charge, and chirality.

In another aspect, the invention is directed to a computer-implemented method for determining whether properties of a query sequence of chemical units match properties of a polymer of chemical units. The query sequence may be represented by a first data structure, tangibly embodied in a computer-readable medium, including an identifier that may include one or more bit fields for storing values corresponding to properties of the query sequence. The polymer may be represented by a second data structure, tangibly embodied in a computer-readable medium, including an identifier that may include one or more bit fields for storing values corresponding to properties of the polymer. The method may include acts of generating at least one mask based on the values stored in the one or more bit fields of the first data structure, performing at least one binary operation on the values stored in the one or more bit fields of the second data structure using the at least one mask to generate at least one result, and determining whether the properties of the query sequence match the properties of the polymer based on the at least one result. The chemical units may, for example, be any of the chemical units described above. Similarly, the properties may be any of the properties described above.

In one embodiment, the act of generating includes an act of generating the at least one mask as a sequence of bits that is equivalent to the values stored in the one or more bit fields of the first data structure. In another embodiment, the act of generating includes an act of generating the at least one mask as a sequential repetition of the values stored in the one or more bit fields of the first data structure.

In a further embodiment, the at least one mask includes a plurality of masks and the act of performing at least one binary operation includes acts of performing a logical AND operation on the values stored in the one or more bit fields of the second data structure using each of the plurality of masks to generate a plurality of intermediate results, and combining the plurality of intermediate results using at least one logical OR operation to generate the at least one result. In one embodiment, the act of determining includes an act of determining that the properties of the query sequence match the properties of the polymer when the at least one result has a non-zero value. In a further embodiment, the at least one binary operation includes at least one logical AND operation.

In another aspect, the invention is directed to a database, tangibly embodied in a computer-readable medium, for storing information descriptive of one or more polymers. The database may include one or more data units (e.g., records or table entries) corresponding to the one or more polymers, each of the data units may include an identifier

that may include one or more fields for storing values corresponding to properties of the polymer.

In another embodiment, the invention is directed to a data structure, tangibly embodied in a computer-readable medium, representing a chemical unit of a polymer. The data structure may comprise an identifier including one or more fields. Each field may be for storing a value corresponding to one or more properties of the chemical unit. At least one field may store a non-character-based value such as, for example, a binary or decimal value.

Other aspects of the invention include the various combinations of one or more of the foregoing aspects of the invention, as well as the combinations of one or more of the various embodiments thereof as found in the following detailed description or as may be derived therefrom. It should be understood that the foregoing aspects of the invention also have corresponding computer-implemented processes which are also aspects of the present invention. It should also be understood that other embodiments of the present invention may be derived by those of ordinary skill in the art both from the following detailed description of a particular embodiment of the invention.

Brief Description of the Drawings

FIG. 1 is a block diagram illustrating an example of a computer system for storing and manipulating polymer information.

FIG. 2A is a diagram illustrating an example of a record for storing information about a polymer and its constituent chemical units.

FIG. 2B is a diagram illustrating an example of a record for storing information about a polymer.

FIG. 2C is a diagram illustrating an example of a record for storing information about constituent chemical units of a polymer.

FIG. 3 is a flow chart illustrating an example of a method for determining whether properties of a first polymer of chemical units match properties of a second chemical unit.

Detailed Description

The present invention will be better understood in view of the following detailed description of a particular embodiment thereof, taken in conjunction with the attached drawings. All references cited herein are hereby expressly incorporated by reference.

FIG. 1 shows an example of a computer system 100 for storing and manipulating polymer information. The computer system 100 includes a polymer database 102 which

includes a plurality of records 104a-n storing information corresponding to a plurality of polymers. Each of the records 104a-n may store information about properties of the corresponding polymer, properties of the corresponding polymer's constituent chemical units, or both. The polymers for which information is stored in the polymer database 102
5 may be any kind of polymers. For example, the polymers may include polysaccharides, nucleic acids, or polypeptides.

A "polymer" as used herein is a compound having a linear and/or branched backbone of chemical units which are secured together by linkages. In some but not all cases the backbone of the polymer may be branched. The term "backbone" is given its
10 usual meaning in the field of polymer chemistry. The polymers may be heterogeneous in backbone composition thereby containing any possible combination of polymer units linked together such as peptide- nucleic acids. In an embodiment, a polymer is homogeneous in backbone composition and is, for example, a nucleic acid, a polypeptide, a polysaccharide, a carbohydrate, a polyurethane, a polycarbonate, a polyurea, a polyethyleneimine, a
15 polyarylene sulfide, a polysiloxane, a polyimide, a polyacetate, a polyamide, a polyester, or a polythioester. A "polysaccharide" is a biopolymer comprised of linked saccharide or sugar units. A "nucleic acid" as used herein is a biopolymer comprised of nucleotides, such as deoxyribose nucleic acid (DNA) or ribose nucleic acid (RNA). A polypeptide as used herein is a biopolymer comprised of linked amino acids.

As used herein with respect to linked units of a polymer, "linked" or "linkage"
20 means two entities are bound to one another by any physicochemical means. Any linkage known to those of ordinary skill in the art, covalent or non-covalent, is embraced. Such linkages are well known to those of ordinary skill in the art. Natural linkages, which are those ordinarily found in nature connecting the chemical units of a particular polymer, are
25 most common. Natural linkages include, for instance, amide, ester and thioester linkages. The chemical units of a polymer analyzed by the methods of the invention may be linked, however, by synthetic or modified linkages. Polymers where the units are linked by covalent bonds will be most common but also include hydrogen bonded, etc.

The polymer is made up of a plurality of chemical units. A "chemical unit" as used
30 herein is a building block or monomer which can be linked directly or indirectly to other building blocks or monomers to form a polymer. The polymer preferably is a polymer of at least two different linked units. The particular type of unit will depend on the type of polymer. For instance DNA is a biopolymer comprised of a deoxyribose phosphate backbone composed of units of purines and pyrimidines such as adenine, cytosine, guanine,

thymine, 5-methylcytosine, 2-aminopurine, 2-amino-6-chloropurine, 2,6-diaminopurine, hypoxanthine, and other naturally and non-naturally occurring nucleobases, substituted and unsubstituted aromatic moieties. RNA is a biopolymer comprised of a ribose phosphate backbone composed of units of purines and pyrimidines such as those described for DNA but wherein uracil is substituted for thymidine. DNA units may be linked to the other units of the polymer by their 5' or 3' hydroxyl group thereby forming an ester linkage. RNA units may be linked to the other units of the polymer by their 5', 3' or 2' hydroxyl group thereby forming an ester linkage. Alternatively, DNA or RNA units having a terminal 5', 3' or 2' amino group may be linked to the other units of the polymer by the amino group thereby forming an amide linkage.

Whenever a nucleic acid is represented by a sequence of letters it will be understood that the nucleotides are in 5'→3' order from left to right and that "A" denotes adenosine, "C" denotes cytidine, "G" denotes guanosine, "T" denotes thymidine, and "U" denotes uracil unless otherwise noted.

The chemical units of a polypeptide are amino acids, including the 20 naturally occurring amino acids as well as modified amino acids. Amino acids may exist as amides or free acids and are linked to the other units in the backbone of the polymers through their α-amino group thereby forming an amide linkage to the polymer.

A polysaccharide is a polymer composed of monosaccharides linked to one another. In many polysaccharides the basic building block of the polysaccharide is actually a disaccharide unit which can be repeating or non-repeating. Thus, a unit when used with respect to a polysaccharide refers to a basic building block of a polysaccharide and can include a monomeric building block (monosaccharide) or a dimeric building block (disaccharide).

A "plurality of chemical units" is at least two units linked to one another.

The polymers may be native or naturally-occurring polymers which occur in nature or non-naturally occurring polymers which do not exist in nature. The polymers typically include at least a portion of a naturally occurring polymer. The polymers can be isolated or synthesized *de novo*. For example, the polymers can be isolated from natural sources e.g. purified, as by cleavage and gel separation or may be synthesized e.g., (i) amplified *in vitro* by, for example, polymerase chain reaction (PCR); (ii) synthesized by, for example, chemical synthesis; (iii) recombinantly produced by cloning, etc.

Fig. 2A illustrates an example of the format of a data unit 200 in the polymer database 102 (i.e., one of the data units 104a-n). As shown in FIG. 2A, the data unit 200

may include a polymer identifier (ID) 202 that identifies the polymer corresponding to the data unit 200. The polymer ID 202 is described in more detail below with respect to FIG. 2B. The data unit 200 also may include one or more chemical unit identifiers (IDs) 204a-*n* corresponding to chemical units that are constituents of the polymer corresponding to the data unit 200. The chemical unit IDs 204a-*n* are described in more detail below with respect to FIG. 2C. The format of the data unit 200 shown in FIG. 2A is merely an example of a format that may be used to represent polymers in the polymer database 102. Polymers may be represented in the polymer database in other ways. For example, the data unit 200 may include only the polymer ID 202 or may only include one or more of the chemical unit IDs 204a-*n*.

FIG. 2B illustrates an example of the polymer ID 202. The polymer ID 202 may include one or more fields 202a-*n* for storing information about properties of the polymer corresponding to the data unit 200 (FIG. 2A). Similarly, FIG. 2C illustrates an example of the chemical unit 204a. The chemical unit ID 204a may include one or more fields 206a-*m* for storing information about properties of the chemical unit corresponding to the chemical unit ID 204a. Although the following description refers to the fields 206a-*m* of the chemical unit ID 204a, such description is equally applicable to the fields 202a-*n* of the polymer ID 202a (and the fields of the chemical unit IDs 204b-*n*).

The fields 206a-*m* of the chemical unit ID 204a may store any kind of value that is capable of being stored in a computer readable medium, such as, for example, a binary value, a hexadecimal value, an integral decimal value, or a floating point value.

Each field 206a-*m* may store information about any property of the corresponding chemical unit. A "property" as used herein is a characteristic (e.g., structural characteristic) of the polymer that provides information (e.g., structural information) about the polymer.

When the term property is used with respect to any polymer except a polysaccharide the property provides information other than the identity of a unit of the polymer or the polymer itself. A compilation of several properties of a polymer may provide sufficient information to identify a chemical unit or even the entire polymer but the property of the polymer itself does not encompass the chemical basis of the chemical unit or polymer.

When the term property is used with respect to polysaccharides, to define a polysaccharide property, it has the same meaning as described above except that due to the complexity of the polysaccharide, a property may identify a type of monomeric building block of the polysaccharide. Chemical units of polysaccharides are much more complex than chemical units of other polymers, such as nucleic acids and polypeptides. The

polysaccharide unit has more variables in addition to its basic chemical structure than other chemical units. For example, the polysaccharide may be acetylated or sulfated at several sites on the chemical unit, or it may be charged or uncharged. Thus, one property of a polysaccharide may be the identity of one or more basic building blocks of the

5 polysaccharides.

A basic building block alone, however, may not provide information about the charge and the nature of substituents of the saccharide or disaccharide. For example, a building block of uronic acid may be iduronic or glucuronic acid. Each of these building blocks may have additional substituents that add complexity to the structure of the chemical

10 unit. A single property, however, may not identify such additional substitutes charges, etc., in addition to identifying a complete building block of a polysaccharide. This information, however, may be assembled from several properties. Thus, a property of a polymer as used herein does not encompass an amino acid or nucleotide but does encompass a saccharide or disaccharide building block of a polysaccharide.

15 A type of property that provides information about a polymer may depend on a type of polymer being analyzed. For instance, if the polymer is a polysaccharide, properties such as charge, molecular weight, nature and degree of sulfation or acetylation, and type of saccharide may provide information about the polymer. Properties may include, but are not limited to, charge, chirality, nature of substituents, quantity of substituents, molecular

20 weight, molecular length, compositional ratios of substituents or units, type of basic building block of a polysaccharide, hydrophobicity, enzymatic sensitivity, hydrophilicity, secondary structure and conformation (i.e., position of helices), spatial distribution of substituents, ratio of one set of modifications to another set of modifications (i.e., relative amounts of 2-O sulfation to N-sulfation or ratio of iduronic acid to glucuronic acid), and binding sites for

25 proteins. Other properties may be identified by those of ordinary skill in the art. A substituent, as used herein is an atom or group of atoms that substitute a unit, but are not themselves the units.

A property of a polymer may be identified by any means known in the art. The procedure used to identify a property may depend on a type of property. Molecular weight,

30 for instance, may be determined by several methods including mass spectrometry. The use of mass spectrometry for determining the molecular weight of polymers is well known in the art. Mass Spectrometry has been used as a powerful tool to characterize polymers because of its accuracy (± 1 Dalton) in reporting the masses of fragments generated (e.g., by enzymatic cleavage), and also because only pM sample concentrations are required. For

example, matrix-assisted laser desorption ionization mass spectrometry (MALDI-MS) has been described for identifying the molecular weight of polysaccharide fragments in publications such as Rhomberg, A. J. et al, *PNAS, USA*, v. 95, p. 4176-4181 (1998); Rhomberg, A. J. et al, *PNAS, USA*, v. 95, p. 12232-12237 (1998); and Ernst, S. et. al., *PNAS, USA*, v. 95, p. 4182-4187 (1998), each of which is hereby incorporated by reference. Other types of mass spectrometry known in the art, such as, electron spray-MS, fast atom bombardment mass spectrometry (FAB-MS) and collision-activated dissociation mass spectrometry (CAD) can also be used to identify the molecular weight of the polymer or polymer fragments.

The mass spectrometry data may be a valuable tool to ascertain information about the polymer fragment sizes after the polymer has undergone degradation with enzymes or chemicals. After a molecular weight of a polymer is identified, it may be compared to molecular weights of other known polymers. Because masses obtained from the mass spectrometry data are accurate to one Dalton (1D), a size of one or more polymer fragments obtained by enzymatic digestion may be precisely determined, and a number of substituents (i.e., sulfates and acetate groups present) may be determined. One technique for comparing molecular weights is to generate a mass line and compare the molecular weight of the unknown polymer to the mass line to determine a subpopulation of polymers which have the same molecular weight. A "mass line" as used herein is an information database, preferably in the form of a graph or chart which stores information for each possible type of polymer having a unique sequence based on the molecular weight of the polymer. Thus, a mass line may describe a number of polymers having a particular molecular weight. A two-unit nucleic acid molecule (i.e., a nucleic acid having two chemical units) has 16 (4 units^2) possible polymers at a molecular weight corresponding to two nucleotides. A two-unit polysaccharide (i.e., disaccharide) has 32 possible polymers at a molecular weight corresponding to two saccharides. Thus, a mass line may be generated by uniquely assigning a particular mass to a particular length of a given fragment (all possible di, tetra, hexa, octa, up to a hexadecasaccharide), and tabulating the results (An Example is shown in Figure 4).

Table 1 below shows an example of a computed set of values for a polysaccharide. From Table 1, a number of chemical units of a polymer may be determined from the minimum difference in mass between a fragment of length $n+1$ and a fragment of length n . For example, if the repeat is a disaccharide unit, a fragment of length n has $2n$

monosaccharide units. For example, $n=1$ may correspond to a length of a disaccharide and $n=2$ may correspond to a length of a tetrasaccharide, etc.

| Fragment Length n | Minimum difference in mass between $n+1$ and n (D) |
|---------------------|--|
| 1 | 101.13 |
| 2 | 13.03 |
| 3 | 13.03 |
| 4 | 9.01 |
| 5 | 9.01 |
| 6 | 4.99 |
| 7 | 4.99 |
| 8 | 0.97 |
| 9 | 0.97 |

TABLE 1

Because mass spectrometry data indicates the mass of a fragment to 1D accuracy, a length may be assigned uniquely to fragment by looking up a mass on the mass line.

Further, it may be determined from the mass line that, within a fragment of particular length higher than a disaccharide, there is a minimum of 4.02D different in masses indicating that two acetate groups (84.08D) replaced a sulfate group (80.06D). Therefore, a number of sulfates and acetates of a polymer fragment may be determined from the mass from the mass spectrometry data and, such number may be assigned to the polymer fragment.

In addition to molecular weight, other properties may be determined using methods known in the art. The compositional ratios of substituents or chemical units (quantity and type of total substituents or chemical units) may be determined using methodology known in the art, such as capillary electrophoresis. A polymer may be subjected to an experimental constraint such as enzymatic or chemical degradation to separate each of the chemical units of the polymers. These units then may be separated using capillary electrophoresis to determine the quantity and type of substituents or chemical units present in the polymer. Additionally, a number of substituents or chemical units can be determined using calculations based on the molecular weight of the polymer.

In the method of capillary gel-electrophoresis, reaction samples may be analyzed by small-diameter, gel-filled capillaries. The small diameter of the capillaries (50 μm) allows for efficient dissipation of heat generated during electrophoresis. Thus, high field strengths can be used without excessive Joule heating (400 V/m), lowering the separation time to about 20 minutes per reaction run, therefor increasing resolution over conventional gel electrophoresis. Additionally, many capillaries may be analyzed in parallel, allowing amplification of generated polymer information.

In addition to being useful for identifying a property, compositional analysis also may be used to determine a presence and composition of an impurity as well as a main property of the polymer. Such determinations may be accomplished if the impurity does not contain an identical composition as the polymer. To determine whether an impurity is present may involve accurately integrating an area under each peak that appears in the electrophoretogram and normalizing the peaks to the smallest of the major peaks. The sum of the normalized peaks should be equal to one or close to being equal to one. If it is not, then one or more impurities are present. Impurities even may be detected in unknown samples if at least one of the disaccharide units of the impurity differs from any disaccharide unit of the unknown.

If an impurity is present, one or more aspects of a composition of the components may be determined using capillary electrophoresis. Because all known disaccharide units may be baseline-separated by the capillary electrophoresis method described above and because migration times typically are determined using electrophoresis (i.e., as opposed to electroosmotic flow) and are reproducible, reliable assignment to a polymer fragment of the various saccharide units may be achieved. Consequently, both a composition of the major peak and a composition of a minor contaminant may be assigned to a polymer fragment. The composition for both the major and minor components of a solution may be assigned as described below.

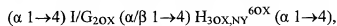
One example of such assignment of compositions involves determining the composition of the major AT-III binding HLGAG deca-saccharide (+ DDD4-7) and its minor contaminant (+ D5D4-7) present in solution in a 9:1 ratio. Complete digestion of this 9:1 mixture with a heparinases yields 4 peaks: three representative of the major deca-saccharide (viz., D, 4, and -7) which are also present in the contaminant and one peak, 5, that is present only in the contaminant. In other words, the area of each peak for D, 4, and -7 represents an additive combination of a contribution from the major deca-saccharide and

the contribution from the contaminant, whereas the peak for 5 represents only the contaminant.

To assign the composition of the contaminant and the major component, the area under the 5 peak may be used as a starting point. This area represents an area under the peak for one disaccharide unit of the contaminant. Subtracting this area from the total area of 4 and -7 and subtracted twice this area from an area under D yields a 1:1:3 ratio of 4:-7:D. Such a ratio confirms the composition of the major component and indicates that the composition of the impurity is two Ds, one 4, one -7 and one 5.

Methods of identifying other types of properties may be easily identifiable to those of skill in the art and may depend on the type of property and the type of polymer. For example, hydrophobicity may be determined using reverse-phase high-pressure liquid chromatography (RP-HPLC). Enzymatic sensitivity may be identified by exposing the polymer to an enzyme and determining a number of fragments present after such exposure. The chirality may be determined using circular dichroism. Protein binding sites may be determined by mass spectrometry, isothermal calorimetry and NMR. Enzymatic modification (not degradation) may be determined in a similar manner as enzymatic degradation, i.e., by exposing a substrate to the enzyme and using MALDI-MS to determine if the substrate is modified. For example, a sulfotransferase may transfer a sulfate group to an HS chain having a concomitant increase in 80Da. Conformation may be determined by modeling and nuclear magnetic resonance (NMR). The relative amounts of sulfation may be determined by compositional analysis or approximately determined by raman spectroscopy.

FIG. 2D illustrates an example of the chemical unit ID 204a. The chemical unit ID 204a contains one or more fields 212a-e for storing information about properties of a heparin-like glycosaminoglycan (HLGAG). HLGAGs are complex polysaccharide molecules made up of disaccharide repeat units comprising hexoseamine and glucuronic/iduronic acid that are linked by α/β 1-4 glycosidic linkages. These defining units may be modified by: sulfation at the N, 3-O and 6-O position of the hexoseamine, 2-O sulfation of the uronic acid, and C5 epimerization that converts the glucuronic acid to iduronic acid. The disaccharide unit of HLGAG may be represented as:



where X may be sulfated (-SO₃H) or unsulfated (-H), and Y may be sulfated (-SO₃H) or acetylated (-COCH₃) or, in rare cases, neither sulfated nor acetylated.

The fields 212a-e may store any kinds of values, such as, for example single-bit values, single-digit hexadecimal values, or decimal values. In one embodiment, the chemical unit ID 204a includes each of the following fields: (1) a field 212a for storing a value indicating whether the polymer contains an iduronic or a glucuronic acid (I/G); (2) a field 212b for storing a value indicating whether the 2X position of the iduronic or glucuronic acid is sulfated or unsulfated; (3) a field 212c for storing a value indicating whether the hexoseamine is sulfated or unsulfated; (4) a field 212d indicating whether the 3X position of the hexoseamine is sulfated or unsulfated; and (5) a field 212e indicating whether the NX position of the hexoseamine is sulfated or acetylated. Optionally, each of the fields 212a-e may be represented as a single bit.

Table 2 illustrates an example of a data structure having a plurality of entries, where each entry represents an HLGAG encoded in accordance with Fig. 2D. Bit values for each of the fields 212a-e may be assigned in any known manner. For example, with respect to field 212a (I/G), a value of one may indicate Iduronic and a value of zero may indicate Glucuronic, or vice versa.

| I/G | 2X | 6X | 3X | NX | ALPH CODE | DISACC | MASS (ΔU) |
|-----|----|----|----|----|--------------|---------------------------------------|--------------|
| 0 | 0 | 0 | 0 | 0 | 0 | I-H _N Ac | 379.33 |
| 0 | 0 | 0 | 0 | 1 | 1 | I-H _{NS} | 417.35 |
| 0 | 0 | 0 | 1 | 0 | 2 | I-H _N Ac,3S | 459.39 |
| 0 | 0 | 0 | 1 | 1 | 3 | I-H _{NS} ,3S | 497.41 |
| 0 | 0 | 1 | 0 | 0 | 4 | I-H _N Ac,6S | 459.39 |
| 0 | 0 | 1 | 0 | 1 | 5 | I-H _{NS} ,6S | 497.41 |
| 0 | 0 | 1 | 1 | 0 | 6 | I-H _N Ac,3S,6S | 539.45 |
| 0 | 0 | 1 | 1 | 1 | 7 | I-H _{NS} ,3S,6S | 577.47 |
| 0 | 1 | 0 | 0 | 0 | 8 | I _{2S} -H _N Ac | 459.39 |
| 0 | 1 | 0 | 0 | 1 | 9 | I _{2S} -H _{NS} | 497.41 |
| 0 | 1 | 0 | 1 | 0 | A | I _{2S} -H _N Ac,3S | 539.45 |
| 0 | 1 | 0 | 1 | 1 | B | I _{2S} -H _{NS} ,3S | 577.47 |
| 0 | 1 | 1 | 0 | 0 | C | I _{2S} -H _N Ac,6S | 539.45 |
| 0 | 1 | 1 | 0 | 1 | D | I _{2S} -H _{NS} ,6S | 577.47 |

| I/G | 2X | 6X | 3X | NX | ALPH CODE | DISACC | MASS (ΔU) |
|-----|----|----|----|----|--------------|-------------------------------|------------------------|
| 0 | 1 | 1 | 1 | 0 | E | I_{2S^-} $H_{NAc,3S,6S}$ | 619.51 |
| 0 | 1 | 1 | 1 | 1 | F | $I_{2S^-}H_{NS,3S,6S}$ | 657.53 |
| 1 | 0 | 0 | 0 | 0 | -0 | $G-H_{NAc}$ | 379.33 |
| 1 | 0 | 0 | 0 | 1 | -1 | $G-H_{NS}$ | 417.35 |
| 1 | 0 | 0 | 1 | 0 | -2 | $G-H_{NAc,3S}$ | 459.39 |
| 1 | 0 | 0 | 1 | 1 | -3 | $G-H_{NS,3S}$ | 497.41 |
| 1 | 0 | 1 | 0 | 0 | -4 | $G-H_{NAc,6S}$ | 459.39 |
| 1 | 0 | 1 | 0 | 1 | -5 | $G-H_{NS,6S}$ | 497.41 |
| 1 | 0 | 1 | 1 | 0 | -6 | $G-H_{NAc,3S,6S}$ | 539.45 |
| 1 | 0 | 1 | 1 | 1 | -7 | $G-H_{NS,3S,6S}$ | 577.47 |
| 1 | 1 | 0 | 0 | 0 | -8 | $G_{2S^-}H_{NAc}$ | 459.39 |
| 1 | 1 | 0 | 0 | 1 | -9 | $G_{2S^-}H_{NS}$ | 497.41 |
| 1 | 1 | 0 | 1 | 0 | -A | $G_{2S^-}H_{NAc,3S}$ | 539.45 |
| 1 | 1 | 0 | 1 | 1 | -B | $G_{2S^-}H_{NS,3S}$ | 577.47 |
| 1 | 1 | 1 | 0 | 0 | | $G_{2S^-}H_{NAc,6S}$ | |
| 1 | 1 | 1 | 0 | 1 | -D | $G_{2S^-}H_{NS,6S}$ | 577.47 |
| 1 | 1 | 1 | 1 | 0 | -E | G_{2S^-} $H_{NAc,3S,6S}$ | 619.51 |
| 1 | 1 | 1 | 1 | 1 | -F | G_{2S^-} $H_{NS,3S,6S}$ | 657.53 |

TABLE 2

Representing a HLGAG using a bit field may have a number of advantages.

- 5 Because a property of an HLGAG may have one of two possible states, a binary bit is ideally-suited for storing information representing an HLGAG property. Bit fields may be used to store such information in a computer readable medium (e.g., a computer memory or storage device), for example, by packing multiple bits (representing multiple fields) into a single byte or sequence of bytes. Furthermore, bit fields may be stored and manipulated
- 10 quickly and efficiently by digital computer processors, which typically store information

using bits and which typically can quickly perform operations (e.g., shift, AND, OR) on bits. For example, as described in more detail below, a plurality of properties each stored as a bit field can be searched more quickly than searches conducted using typical character-based searching methods.

5 Further, using bit fields to represent properties of HLGAGs permits a user to more easily incorporate additional properties (e.g., 4-O sulfation vs. unsulfation) into a chemical unit ID 204a by adding extra bits to represent the additional properties.

In one embodiment, the four fields 212b-e (each of which may store a single-bit value) may be represented as a single hexadecimal (base 16) number where each of the
10 fields 212a-e represents one bit of the hexadecimal number. Using hexadecimal numbers to represent disaccharide units is convenient both for representation and processing because hexadecimal digits are a common form of representation used by conventional computers.

Optionally, the five fields 212a-e of the record 210 may be represented as signed hexadecimal digit, in which the fields 212b-212e collectively encode a single-digit
15 hexadecimal number as described above and the I/G field is used as a sign bit. In such a signed representation, the hexadecimal numbers 0-F may be used to code chemical units containing iduronic acid and the hexadecimal numbers -0 to -F may be used to code units containing glucuronic acid. The chemical unit ID 204a may, however, be encoded using other forms of representations, such as by using a twos-complement representation.

20 The fields 212a-e of the chemical unit ID 204a may be arranged in any order. For example, a gray code system may be used to code HLGAGs. In a gray code numbering scheme, each successive value differs from the previous value only in a single bit position. For example, in the case of HLGAGs, the values representing HLGAGs may be arranged so that any two neighboring values differ in the value of only one property. An example of a
25 gray code system used to code HLGAGs is shown in Table 3.

| I/G | 2X | 6X | 3X | NX | Numeric | DISACC | MASS |
|-----|----|----|----|----|---------|---------------------------|--------|
| 16 | 8 | 4 | 2 | 1 | Value | | (ΔU) |
| 0 | 0 | 0 | 0 | 0 | 0 | I-H _N Ac | 379.33 |
| 0 | 0 | 0 | 0 | 1 | 1 | I-H _N S | 417.35 |
| 0 | 0 | 0 | 1 | 1 | 3 | I-H _N S,3S | 497.41 |
| 0 | 0 | 0 | 1 | 0 | 2 | I-H _N Ac,3S | 459.39 |
| 0 | 0 | 1 | 1 | 0 | 6 | I-H _N Ac,3S,6S | 539.45 |

| I/G | 2X | 6X | 3X | NX | Numeric Value | DISACC | MASS (ΔU) |
|-----|----|----|----|----|------------------|---|------------------------|
| 16 | 8 | 4 | 2 | 1 | | | |
| 0 | 0 | 1 | 1 | 1 | 7 | I-H _{NS,3S,6S} | 577.47 |
| 0 | 0 | 1 | 0 | 1 | 5 | I-H _{NS,6S} | 497.41 |
| 0 | 0 | 1 | 0 | 0 | 4 | I-H _{NAC,6S} | 459.39 |
| 0 | 1 | 1 | 0 | 0 | 12 | I _{2S} -H _{NAC,6S} | 539.45 |
| 0 | 1 | 1 | 0 | 1 | 13 | I _{2S} -H _{NS,6S} | 577.47 |
| 0 | 1 | 1 | 1 | 1 | 15 | I _{2S} -H _{NS,3S,6S} | 657.53 |
| 0 | 1 | 1 | 1 | 0 | 14 | I _{2S} -H _{NAC,3S,6S} | 619.51 |
| 0 | 1 | 0 | 1 | 0 | 10 | I _{2S} -H _{NAC,3S} | 539.45 |
| 0 | 1 | 0 | 1 | 1 | 11 | I _{2S} -H _{NS,3S} | 577.47 |
| 0 | 1 | 0 | 0 | 1 | 9 | I _{2S} -H _{NS} | 497.41 |
| 0 | 1 | 0 | 0 | 0 | 8 | I _{2S} -H _{NAC} | 459.39 |
| 1 | 1 | 0 | 0 | 0 | 24 | G _{2S} -H _{NAC} | 459.39 |
| 1 | 1 | 0 | 0 | 1 | 25 | G _{2S} -H _{NS} | 497.41 |
| 1 | 1 | 0 | 1 | 1 | 27 | G _{2S} -H _{NS,3S} | 577.41 |
| 1 | 1 | 0 | 1 | 0 | 26 | G _{2S} -H _{NAC,3S} | 539.45 |
| 1 | 1 | 1 | 1 | 0 | 30 | G _{2S} -H _{NAC,3S,6S} | 619.51 |
| 1 | 1 | 1 | 1 | 1 | 31 | G _{2S} -H _{NS,3S,6S} | 657.53 |
| 1 | 1 | 1 | 0 | 1 | 29 | G _{2S} -H _{NS,6S} | 577.47 |
| 1 | 1 | 1 | 0 | 0 | 28 | G _{2S} -H _{NAC,6S} | 539.45 |
| 1 | 0 | 1 | 0 | 0 | 20 | G-H _{NAC,6S} | 459.39 |
| 1 | 0 | 1 | 0 | 1 | 21 | G-H _{NS,6S} | 497.41 |
| 1 | 0 | 1 | 1 | 1 | 23 | G-H _{NS,3S,6S} | 577.47 |
| 1 | 0 | 1 | 1 | 0 | 22 | G-H _{NAC,3S,6S} | 539.45 |
| 1 | 0 | 0 | 1 | 0 | 18 | G-H _{NAC,3S} | 459.39 |
| 1 | 0 | 0 | 1 | 1 | 19 | G-H _{NS,3S} | 497.41 |
| 1 | 0 | 0 | 0 | 1 | 17 | G-H _{NS} | 417.35 |
| 1 | 0 | 0 | 0 | 0 | 16 | G-H _{NAC} | 379.33 |

TABLE 3

Table 3 illustrates that use of a gray coding scheme arranges the disaccharide
 5 building blocks such that neighboring table entries differ from each other only in the value

of a single property. One advantage of using gray codes to encode HLGAGs is that a biosynthesis of HLGAG fragments may follow a specific sequence of modifications starting from the basic building block G-H_{NAc}.

In Table 3, bit weights of 8, 4, 2, and 1 are used to calculate the numerical equivalent of a hexadecimal number with the most significant bit (I/G) being used as a sign bit. For example, the hexadecimal code A (01010 binary) is equal to $8*1 + 4*0 + 2*1 + 1*0 = 10$.

In another embodiment, the weights of each of the fields 212a-e may be changed thereby implementing an alternative weighting system. For example, bit fields 212a-e may have weights of 16, 8, 4, -2, and -1, respectively, as shown in Table 4.

10

| I/G | 2X | NX | 3X | 6X | Value | DISACC | MASS (ΔU) |
|-----|----|----|----|----|-------|---|--------------|
| 16 | 8 | 4 | -2 | -1 | | | |
| 0 | 0 | 0 | 0 | 0 | 0 | I-H _{NAc} | 379.33 |
| 0 | 0 | 0 | 0 | 1 | -1 | I-H _{NAc,6S} | 459.39 |
| 0 | 0 | 0 | 1 | 0 | -2 | I-H _{NAc,3S} | 459.39 |
| 0 | 0 | 0 | 1 | 1 | -3 | I-H _{NAc,3S,6S} | 539.45 |
| 0 | 0 | 1 | 0 | 0 | 4 | I-H _{NS} | 417.35 |
| 0 | 0 | 1 | 0 | 1 | 3 | I-H _{NS,6S} | 497.41 |
| 0 | 0 | 1 | 1 | 0 | 2 | I-H _{NS,3S} | 497.41 |
| 0 | 0 | 1 | 1 | 1 | 1 | I-H _{NS,3S,6S} | 577.47 |
| 0 | 1 | 0 | 0 | 0 | 8 | I _{2S} -H _{NAc} | 459.39 |
| 0 | 1 | 0 | 0 | 1 | 7 | I _{2S} -H _{NAc,6S} | 539.45 |
| 0 | 1 | 0 | 1 | 0 | 6 | I _{2S} -H _{NAc,3S} | 539.45 |
| 0 | 1 | 0 | 1 | 1 | 5 | I _{2S} -H _{NAc,3S,6S} | 619.51 |
| 0 | 1 | 1 | 0 | 0 | 12 | I _{2S} -H _{NS} | 497.41 |
| 0 | 1 | 1 | 0 | 1 | 11 | I _{2S} -H _{NS,6S} | 577.47 |
| 0 | 1 | 1 | 1 | 0 | 10 | I _{2S} -H _{NS,3S} | 577.47 |
| 0 | 1 | 1 | 1 | 1 | 9 | I _{2S} -H _{NS,3S,6S} | 657.53 |
| 1 | 0 | 0 | 0 | 0 | 16 | G-H _{NAc} | 379.33 |
| 1 | 0 | 0 | 0 | 1 | 15 | G-H _{NAc,6S} | 459.39 |
| 1 | 0 | 0 | 1 | 0 | 14 | G-H _{NAc,3S} | 459.39 |
| 1 | 0 | 0 | 1 | 1 | 13 | G-H _{NAc,3S,6S} | 539.45 |

| I/G | 2X | NX | 3X | 6X | Value | DISACC | MASS (ΔU) |
|-----|----|----|----|----|-------|---|--------------|
| 16 | 8 | 4 | -2 | -1 | | | |
| 1 | 0 | 1 | 0 | 0 | 20 | G-H _{NS} | 417.35 |
| 1 | 0 | 1 | 0 | 1 | 19 | G-H _{NS,6S} | 497.41 |
| 1 | 0 | 1 | 1 | 0 | 18 | G-H _{NS,3S} | 497.41 |
| 1 | 0 | 1 | 1 | 1 | 17 | G-H _{NS,3S,6S} | 577.47 |
| 1 | 1 | 0 | 0 | 0 | 24 | G _{2S} -H _{NAC} | 459.39 |
| 1 | 1 | 0 | 0 | 1 | 23 | G _{2S} -H _{NAC,6S} | 539.45 |
| 1 | 1 | 0 | 1 | 0 | 22 | G _{2S} -H _{NAC,3S} | 539.45 |
| 1 | 1 | 0 | 1 | 1 | 21 | G _{2S} -H _{NAC,3S,6S} | 619.51 |
| 1 | 1 | 1 | 0 | 0 | 28 | G _{2S} -H _{NS} | 497.41 |
| 1 | 1 | 1 | 0 | 1 | 27 | G _{2S} -H _{NS,6S} | 577.47 |
| 1 | 1 | 1 | 1 | 0 | 26 | G _{2S} -H _{NS,3S} | 577.47 |
| 1 | 1 | 1 | 1 | 1 | 25 | G _{2S} -H _{NS,3S,6S} | 657.53 |

TABLE 4

Modifying the weights of the bits may be used to score the disaccharide units. For example, a database of sequences may be created and the different disaccharide units may be scored based on their relative abundance in the sequences present in the database. Some units, for example, I-H_{NAC,3S}^{6S}, which rarely occur in naturally-occurring HLGAGs, may receive a low score based on a scheme in which the bits are weighted in the manner shown in Table 4.

Optionally, the sulfation and acetylation positions may be arranged in an shown in Table 2: I/G, 2X, 6X, 3X, NX. These positions may, however, be arranged differently, resulting in a same set of codes representing different disaccharide units. Table 5, for example, shows an arrangement in which the positions are arranged as I/G, 2X, NX, 3X, 6X.

| I/G | 2X | NX | 3X | 6X | ALPH CODE | DISACC | MASS (ΔU) |
|-----|----|----|----|----|--------------|-----------------------|--------------|
| 0 | 0 | 0 | 0 | 0 | 0 | I-H _{NAC} | 379.33 |
| 0 | 0 | 0 | 0 | 1 | 1 | I-H _{NAC,6S} | 459.39 |
| 0 | 0 | 0 | 1 | 0 | 2 | I-H _{NAC,3S} | 459.39 |

| I/G | 2X | NX | 3X | 6X | ALPH CODE | DISACC | MASS (Δ U) |
|-----|----|----|----|----|--------------|---|-----------------------|
| 0 | 0 | 0 | 1 | 1 | 3 | I-H _{NAC,3S,6S} | 539.45 |
| 0 | 0 | 1 | 0 | 0 | 4 | I-H _{NS} | 417.35 |
| 0 | 0 | 1 | 0 | 1 | 5 | I-H _{NS,6S} | 497.41 |
| 0 | 0 | 1 | 1 | 0 | 6 | I-H _{NS,3S} | 497.41 |
| 0 | 0 | 1 | 1 | 1 | 7 | I-H _{NS,3S,6S} | 577.47 |
| 0 | 1 | 0 | 0 | 0 | 8 | I _{2S} -H _{NAC} | 459.39 |
| 0 | 1 | 0 | 0 | 1 | 9 | I _{2S} -H _{NAC,6S} | 539.45 |
| 0 | 1 | 0 | 1 | 0 | A | I _{2S} -H _{NAC,3S} | 539.45 |
| 0 | 1 | 0 | 1 | 1 | B | I _{2S} - H _{NAC,3S,6S} | 619.51 |
| 0 | 1 | 1 | 0 | 0 | C | I _{2S} -H _{NS} | 497.41 |
| 0 | 1 | 1 | 0 | 1 | D | I _{2S} -H _{NS,6S} | 577.47 |
| 0 | 1 | 1 | 1 | 0 | E | I _{2S} -H _{NS,3S} | 577.47 |
| 0 | 1 | 1 | 1 | 1 | F | I _{2S} -H _{NS,3S,6S} | 657.53 |
| 1 | 0 | 0 | 0 | 0 | -0 | G-H _{NAC} | 379.33 |
| 1 | 0 | 0 | 0 | 1 | -1 | G-H _{NAC,6S} | 459.39 |
| 1 | 0 | 0 | 1 | 0 | -2 | G-H _{NAC,3S} | 459.39 |
| 1 | 0 | 0 | 1 | 1 | -3 | G-H _{NAC,3S,6S} | 539.45 |
| 1 | 0 | 1 | 0 | 0 | -4 | G-H _{NS} | 417.35 |
| 1 | 0 | 1 | 0 | 1 | -5 | G-H _{NS,6S} | 497.41 |
| 1 | 0 | 1 | 1 | 0 | -6 | G-H _{NS,3S} | 497.41 |
| 1 | 0 | 1 | 1 | 1 | -7 | G-H _{NS,3S,6S} | 577.47 |
| 1 | 1 | 0 | 0 | 0 | -8 | G _{2S} -H _{NAC} | 459.39 |
| 1 | 1 | 0 | 0 | 1 | -9 | G _{2S} -H _{NAC,6S} | 539.45 |
| 1 | 1 | 0 | 1 | 0 | -A | G _{2S} -H _{NAC,3S} | 539.45 |
| 1 | 1 | 0 | 1 | 1 | -B | G _{2S} - H _{NAC,3S,6S} | 619.51 |
| 1 | 1 | 1 | 0 | 0 | -C | G _{2S} -H _{NS} | 497.41 |
| 1 | 1 | 1 | 0 | 1 | -D | G _{2S} -H _{NS,6S} | 577.47 |

| I/G | 2X | NX | 3X | 6X | ALPH CODE | DISACC | MASS (ΔU) |
|-----|----|----|----|----|--------------|--|--------------|
| 1 | 1 | 1 | 1 | 0 | -E | G _{2S} -H _{NS,3S} | 577.47 |
| 1 | 1 | 1 | 1 | 1 | -F | G _{2S} - H _{NS,3S,6S} | 657.53 |

TABLE 5

It has been observed that disaccharide units in some HLGAG sequences are neither N-sulfated nor N-acetylated. Such disaccharide units may be represented using the chemical unit ID 204a in any of a number of ways.

If the properties of a chemical unit are represented by bit fields, disaccharide units that contain a free amine in the N position may be represented by, for example, adding an additional bit field. For example, referring to FIG. 2D, an additional field NY may be used in the chemical unit ID 204a. For example, an NY field having a value of zero may correspond to a free amine, and an NY field having a value of one may correspond to N-acetylation, or vice versa. Further, a value of one in the NX field 212e may correspond to N-sulfation.

Optionally, disaccharide units that contain a free amine in the N position may be represented using a tristate field. For example, the field 212e (NX) in the chemical unit ID 204a may be a tristate field having three permissible values. For example, a value of zero may correspond to a free amine, a value of one may correspond to N-acetylation, and a value of two could correspond to N-sulfation. Similarly, the values of any of the fields 212a-e may be represented using a number system with a base higher than two. For example, if the value of the field 212e (NX) is represented by a single-digit number having a base of three, then the field 212e may store three permissible values.

Referring to Fig. 1, user may perform a query on the polymer database 102 to search for particular information. For example, a user may search the polymer database 102 for specified polymers, specified chemical units, or polymers or chemical units having specified properties. A user may provide to a query user interface 108 user input 106 indicating properties for which to search. The user input 106 may, for example, indicate one or more chemical units, a polymer of chemical units or one or more properties to search for using, for example, a standard character-based notation. The query user interface 108 may, for example, provide a graphical user interface (GUI) which allows the user to select from a list of properties using an input device such as a keyboard or a mouse.

The query user interface 108 may generate a search query 110 based on the user input 106. A search engine 112 may receive the search query 110 and generate a mask 114 based on the search query. Example formats of the mask 114, and example techniques to determine whether properties specified by the mask 114 match properties of polymers in the polymer database 102 are described in more detail below in connection to Fig. 3.

The search engine 112 may determine whether properties specified by the mask 114 match properties of polymers stored in the polymer database 102. Subsequently, the search engine 112 may generate search results 116 based on the search indicating whether the polymer database 102 includes polymers having the properties specified by the mask 114.

The search results 116 also may indicate polymers in the polymer database 102 that have the properties specified by the mask 114. For example, if the user input 106 specified properties of a chemical unit, the search results 116 may indicate which polymers in the polymer database 102 include the specified chemical unit. Alternatively, if the user input 106 specified particular chemical unit properties, the search results 116 may indicate polymers in the polymer database 102 that include chemical units having the specified chemical unit properties. Similarly, if the user input 106 specified particular polymer properties, the search results 116 may indicate which polymers in the polymer database 102 have the specified polymer properties.

Fig. 3 is a flowchart illustrating an example of a process 300 that may be used by the search engine 112 to generate the search results 116. In act 302, the search engine 112 may receive a search query 110 from the query user interface 108. Next, in act 304, the search engine 112 may generate a mask 114 generated based on the search query 110. In a following act 306, the search engine 112 may perform a binary operation on one or more of the records 104a-n in the polymer database 102 by applying the mask 114. Next, in act 308, the search engine 112 may generate the search results 116 based on the results of the binary operation performed in step 306.

The process 300 will now be described in more detail with respect to an embodiment in which the fields 206a-m of the chemical unit 204a are binary fields. In act 302, the received search query 110 may indicate to search the polymer database 102 for a particular chemical unit, e.g. the chemical unit I_{2S}-H_{NS}. If, for example, the coding scheme shown in Table 1 is used to encode chemical units in the polymer database, the chemical unit I_{2S}-H_{NS} may be represented by a binary value of 01001. To generate the mask 114 for this chemical unit (step 304), the search engine 112 may use the binary value of the chemical unit, i.e., 01001, as the value of the mask 114. As a result, the values of the bits of the mask 114 may

specify the properties of the chemical unit I_{2S-HNS} . For example, the value of zero in the leftmost bit position may indicate Iduronic, and the value of one in the next bit position may indicate that the 2X position is sulfated.

The search engine 112 may use this mask 114 to determine whether polymers in the polymer database 102 contain the chemical unit I_{2S-HNS} . To make this determination, the search engine 112 may perform a binary operation on the data units $104a-n$ of the polymer database 102 using the mask 114 (step 306). For example, the search engine 112 may perform a logical AND operation on each chemical unit of each of the polymers in the polymer database 102 using the mask 114. If the result of the logical AND operation on a particular chemical unit is equal to the value of the mask 114, then the chemical unit may satisfy the search query 110, and, in act 308, the search engine 112 may indicate a successful match in the search results 116. The search engine 112 may generate additional information in the search results 116, such as the polymer identifier of the polymer containing the matching chemical unit.

In response to receiving the search query in act 302, in act 304, the search engine 112 also may generate the mask 114 that indicates one or more properties of a particular polymer or chemical unit. To generate the mask 114 for such a search query, the search engine 112 may set each bit position in the mask according to a property specified by the search query to the value specified by the search query. Consider, for example, search query 110 that indicates a search for all chemical units in which both the 2X position and the 6X position are sulfated. To generate a mask corresponding to this search query, the search engine 112 may set the bit positions of the mask corresponding to the 2X and 6X positions to a value corresponding to being sulfated. Using the coding scheme shown above in Table 1, for example, in which the 2X and 6X positions have bit positions of 3 and 2 (counting from the rightmost position beginning at bit position zero), respectively, the mask corresponding to this search query is 01100. The two bits of this mask that have a value of one correspond to the bit positions in Table 1 corresponding to the 2X and 6X positions.

To determine whether the one or more properties of a particular chemical unit in the polymer database 102 match the one or more properties specified by the mask 114, the search engine 112 may perform a logical AND operation on the chemical unit identifier of the chemical unit in the polymer database 102 using the mask 114. To generate search results for this chemical unit (i.e., act 308), the search engine 112 may compare the result of the logical AND operation to the mask 114. If the values of the bit positions of the logical AND operation corresponding to the properties specified by the search query are equal to

the values of the same bit positions of the mask 114, then the chemical unit has the properties specified by the search query 110, and the search engine 112 indicates a successful match in the search results 116.

For example, consider the search query 110 described above, which indicates a search for all chemical units in which both the 2X position and the 6X position are sulfated. Using the coding scheme of Table 1, the bit positions corresponding to the 2X and 6X positions are bit positions 3 and 2. Therefore, after performing a logical AND operation on the chemical unit identifier of a chemical unit using the mask 114, the search engine 112 compares bit positions 3 and 2 of the result of the logical AND operation to bit positions 3 and 2 of the mask. If the values in both bit positions are equal, then the chemical unit has the properties specified by the mask 114.

The techniques described above for generating the mask 114 and searching with a mask 114 also may be used to perform searches with respect to sequences of chemical units or entire polymers. For example, if the search query 110 indicates a sequence of chemical units, the search engine 112 may fill the mask 114 with a sequence of bits corresponding to the concatenation of the binary encodings of the specified sequence of chemical units. The search engine 112 may then perform a binary AND operation on the polymer identifiers in the polymer database 102 using the mask 114, and generate the search results 116 as described above.

The techniques described above for generating the mask 114 and searching with the mask 114 are provided merely as an example. Other techniques for generating and searching with the mask 114 may also be used. The search engine 112 also may use more than one mask for each search query 110, and the search engine 112 may perform multiple binary operations in parallel in order to improve computational efficiency. In addition, binary operations other than a logical AND may be used to determine whether properties of the polymers in the polymer database 102 match the properties specified by the mask 114. Other binary operations include, for example, logical OR and logical XOR (exclusive or). Such binary operations may be used alone or in combination with each other.

Using the techniques described above, the polymer database 102 may be searched quickly for particular chemical units. One advantage of the process 300, if used in conjunction with a chemical unit coding scheme that encodes properties of chemical units using binary values is that a chemical unit identifier (e.g., the chemical unit identifier 204a) may be compared to a search query (in the form of a mask) using a single binary operation (e.g., a binary AND operation). As described above, conventional notation systems that use

character-based notation systems to encode sequences of chemical units (e.g., systems which encode DNA sequences as sequences of characters) typically search for a sub-sequence of chemical units (represented by a first sequence of characters) within a super-sequence of chemical units (represented by a second sequence of characters) and use character-based comparison. Such a comparison typically is slow because it sequentially compares each character in a first sequence of characters (corresponding to the sub-sequence) to characters in a second sequence until a match is found. Consequently, the speed of the search is related to the length of the sub-sequence--i.e., the longer the sub-sequence, the slower the search.

In contrast, the speed of the techniques described above for searching binary operations may be constant in relation to the length of a sub-sequence that is the basis for the search query. Because the search engine 112 can search for a query sequence of chemical units using a single binary operation (e.g., a logical AND operation) regardless of the length of the query sequence, searches may be performed more quickly than conventional character-based methods whose speed is related to the length of the query sequence. Further, the binary operations used by the search engine 112 may be performed more quickly because conventional computer processors are designed to perform binary operations on binary data.

A further advantage of the techniques described above for searching using binary operations is that encoding one or more properties of a polymer into the notational representation of the polymer enables the search engine 112 to quickly and directly search the polymer database 102 for particular properties of polymers. Because the properties of a polymer are encoded into the polymer's notational representation, the search engine 112 may determine whether the polymer has a specified property by determining whether the specified property is encoded in the polymer's notational representation. For example, as described above, the search engine 112 may determine whether the polymer has the specified property by performing a logical AND operation on the polymer's notational representation using the mask 114. This operation may be performed quickly by conventional computer processors and may be performed using only the polymer's notational representation and the mask, without reference to additional information about the properties of the polymer.

Some aspects of the techniques described herein for representing properties using binary notation may be useful for generating, searching and manipulating information about polysaccharides. Accordingly, complete building block of a polymer may be assigned a

unique numeric identifier, which may be used to classify the complete building block. For example, each numeric identifier may represent a complete building block of a polysaccharide, including the exact chemical structure as defined by the basic building block of a polysaccharide and all of its substituents, charges etc. A basic building block refers to a basic ring structure such as iduronic acid or glucuronic acid but does not include substituents, charges etc. Such building block information may be generated and processed in a same or similar manner as described above with respect to "properties" of polymers.

A computer system that may implement the system 100 of FIG. 1 as a computer program typically may include a main unit connected to both an output device which displays information to a user and an input device which receives input from a user. The main unit generally includes a processor connected to a memory system via an interconnection mechanism. The input device and output device also may be connected to the processor and memory system via the interconnection mechanism.

One or more output devices may be connected to the computer system. Example output devices include a cathode ray tube (CRT) display, liquid crystal displays (LCD), printers, communication devices such as a modem, and audio output. One or more input devices also may be connected to the computer system. Example input devices include a keyboard, keypad, track ball, mouse, pen and tablet, communication device, and data input devices such as sensors. The subject matter disclosed herein is not limited to the particular input or output devices used in combination with the computer system or to those described herein.

The computer system may be a general purpose computer system which is programmable using a computer programming language, such as C++, Java, or other language, such as a scripting language or assembly language. The computer system also may include specially-programmed, special purpose hardware such as, for example, an Application-Specific Integrated Circuit (ASIC). In a general purpose computer system, the processor typically is a commercially-available processor, of which the series x86, Celeron, and Pentium processors, available from Intel, and similar devices from AMD and Cyrix, the 680X0 series microprocessors available from Motorola, the PowerPC microprocessor from IBM and the Alpha-series processors from Digital Equipment Corporation, are examples. Many other processors are available. Such a microprocessor executes a program called an operating system, of which Windows NT, Linux, UNIX, DOS, VMS and OS8 are examples, which controls the execution of other computer programs and provides scheduling, debugging, input/output control, accounting, compilation, storage assignment, data

management and memory management, and communication control and related services. The processor and operating system define a computer platform for which application programs in high-level programming languages may be written.

A memory system typically includes a computer readable and writeable nonvolatile recording medium, of which a magnetic disk, a flash memory and tape are examples. The disk may be removable, such as a "floppy disk," or permanent, known as a hard drive. A disk has a number of tracks in which signals are stored, typically in binary form, i.e., a form interpreted as a sequence of one and zeros. Such signals may define an application program to be executed by the microprocessor, or information stored on the disk to be processed by the application program. Typically, in operation, the processor causes data to be read from the nonvolatile recording medium into an integrated circuit memory element, which is typically a volatile, random access memory such as a dynamic random access memory (DRAM) or static memory (SRAM). The integrated circuit memory element typically allows for faster access to the information by the processor than does the disk. The processor generally manipulates the data within the integrated circuit memory and then copies the data to the disk after processing is completed. A variety of mechanisms are known for managing data movement between the disk and the integrated circuit memory element, and the subject matter disclosed herein is not limited to such mechanisms. Further, the subject matter disclosed herein is not limited to a particular memory system.

The subject matter disclosed herein is not limited to a particular computer platform, particular processor, or particular high-level programming language. Additionally, the computer system may be a multiprocessor computer system or may include multiple computers connected over a computer network. It should be understood that each module (e.g. 110, 120) in FIG. 1 may be separate modules of a computer program, or may be separate computer programs. Such modules may be operable on separate computers. Data (e.g., 104, 106, 110, 114 and 116) may be stored in a memory system or transmitted between computer systems. The subject matter disclosed herein is not limited to any particular implementation using software or hardware or firmware, or any combination thereof. The various elements of the system, either individually or in combination, may be implemented as a computer program product tangibly embodied in a machine-readable storage device for execution by a computer processor. Various steps of the process may be performed by a computer processor executing a program tangibly embodied on a computer-readable medium to perform functions by operating on input and generating output. Computer programming languages suitable for implementing such a system include

procedural programming languages, object-oriented programming languages, and combinations of the two.

Having now described a few embodiments, it should be apparent to those skilled in the art that the foregoing is merely illustrative and not limiting, having been presented by
5 way of example only. Numerous modifications and other embodiments are within the scope of one of ordinary skill in the art and are contemplated as falling within the scope of the invention.

What is claimed is:

10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
22